音声言語情報処理　9－14
（1995. 12. 15）

テキストコーパスを用いた音声理解のための言語モデル自動獲得

松岡達雄、Robert Hasson、Michael Barlow、古井貞熙

NTTヒューマンインタフェース研究所

連絡先：　松岡達雄
東京都武蔵野市緑町3-9-11
Tel: (0422) 59 3341
E-mail: matsuoka@splab.hil.ntt.jp

あらまし　本報告では音声理解システムにおいて、音声認識結果である自然言語をシステムを駆動する意味言語に変換するための言語モデルを、コーパスから自動的に獲得する方法について述べる。音声理解の問題は自然言語から意味言語へ翻訳する問題ととらえることができる。機械翻訳の研究分野では、二カ国語が一対となった並行テキストからなるコーパスを用いて、翻訳のための言語モデルを統計的に推定する方法が提案されている。本報告ではこれを音声理解における翻訳言語モデルに適用し、スパースデータからの推定の問題を回避するため、統計的手法により文脈自由文法を生成し、有限状態オートマトンで表現された文法ネットワークの状態数を削減することにより、翻訳言語モデルの推定精度を向上する方法を提案する。米国ARPAの音声理解評価タスクである航空旅行情報システム(Air Travel Information System: ATIS)を対象として評価を行い、提案法の有効性を示す。
キーワード：　音声理解、翻訳、言語モデル、自然言語、意味言語

Language model acquisition from a text corpus for speech understanding

Tatsuo Matsuoka, Robert Hasson, Michael Barlow, and Sadaoki Furui

NTT Human Interface Laboratories

Contact:　Tatsuo Matsuoka
3-9-11 Midori-cho, Musashino-shi, Tokyo 180
Tel: (0422) 59 3341
E-mail: matsuoka@splab.hil.ntt.jp

Abstract　Speech understanding can be viewed as a translation problem from natural language into semantic language. This paper describes automatic acquisition of a language model by using a text corpus, which translates natural language into semantic language for speech understanding. A stochastic method for language modeling is proposed for machine translation using a parallel text corpus. This method can be used in speech understanding, but input and output languages should be modeled concisely in order to estimate a reliable translation language model. This paper proposes a method for reducing the number of grammar rules while maintaining the original coverage. This method was shown to be effective by experiments using the ARPA ATIS task.
Keywords: Speech understanding, Translation, Language modeling, Natural language, Semantic language

1. まえがき

我々は米国ARPAの標準評価タスクであるAir Travel Information System（航空旅行情報案内システム、以下ATIS）を対象タスクとして、音声理解の研究を進めている。音声理解を実現するためには、大きく分けて音声入力を自然言語である単語列に変換する音声認識の機能と、単語列から意味を抽出する言語処理の機能が必要である。音声認識に関してはこれまでに、音素コンテキストを考慮した詳細な音響モデルを用いたN-best探索法による方法について報告した[1]。ここでは、音声認識結果の自然言語をデータベース検索言語である意味言語に変換する言語処理に関して報告する。これまで言語処理は、主に人手により文法規則を書いていく方法で実現されていたが、時間・労力の問題、異なるタスクへの移植性の問題などから、自動的に文法規則あるいは言語モデルを生成する方法が望まれている。本報告では統計的手法に基づき、言語モデルをコーパスから自動的に獲得する方法について述べる。

2. 音声理解システムの構成

図1に音声理解システムの構成を示す。音声理解システムは音声認識部と言語処理部からなる。

音声認識部は、Tree-Trellis探索により音声入力に対するN-Best仮説を生成する。N-Best探索に用いる音響モデルは、単語内だけでなく単語間にわたる音素文脈も考慮している[1]。文法としては、語彙単語間の任意の連接を許したno-grammarネットワークを用い、ATISドメインで学習された単語Bigramを言語モデルとして用いる。

言語処理部は、音声認識結果をデータベース検索言語に変換する。ATISタスクにおけるデータベース検索言語はANSI標準のSQLであるが、ATISコーパスには一意にSQL表現に書き直せるWIN(Wizard Input)文が入力文に対応して与えられているので、WIN文を意味言語として実験を行った。本報告では言語処理部に焦点をあてて報告する。

3. 統計的翻訳言語モデル

まず、本報告で提案する方法の基礎となる、機械翻訳のための統計的翻訳言語モデルについて述べる。

自然言語処理研究の分野では、機械による自動翻訳の研究の歴史は古く、1949年にはW. Weaverが統計的手法による機械翻訳を提案している。しかしながら、当時は計算機可読なテキストデータ量が少なく、統計的手法を信頼性高く適用できる状況でなかったことや、計算機の処理性能自体が低かったことなどから発展を見られずに終わった[2,3]。最近になって計算機性能の向上と電子化された文書の増大により、統計的手法が再び盛んに研究されるようになった。

現在の統計的な手法による機械翻訳の基本的枠組みは、Brownらの研究[4-6]による。Brownらは、フランス語から英語への翻訳を行う翻訳言語モデルを

$$P(\mathrm{e}\mid\mathrm{f})=\frac{P(\mathrm{e})P(\mathrm{f}\mid\mathrm{e})}{P(\mathrm{f})} \qquad (1)$$

ただし、e: 英語の文、f: フランス語の文、

として、

$$\hat{\mathrm{e}}=\underset{\mathrm{e}}{\operatorname{argmax}}\ P(\mathrm{e})P(\mathrm{f}\mid\mathrm{e}) \qquad (2)$$

を求めることにより翻訳を行うことを提案した。ここで、言語モデル $P(\mathrm{e})$ は英語のテキストコーパスから推定し、翻訳言語モデル $P(\mathrm{f}\mid\mathrm{e})$ は、フランス語と英語が一対になった並行テキストからなるコーパスを用いて推定する。単にフランス語から英語へ変換するだけならば、直接 $P(\mathrm{e}\mid\mathrm{f})$ を翻訳言語モデルとすればよいように思えるが、翻訳結果の品質は $P(\mathrm{e})P(\mathrm{f}\mid\mathrm{e})$ を用いた方がよくなると述べられている。これは $P(\mathrm{e})$ によってより英語らしい文章が構成されるようになるからであると説明されている[6]。Brownらは、フランス語と英語の並行テキストであるカナダの国会議事録（Hansardコーパス）を用いて、翻訳の実験を行っている。語彙

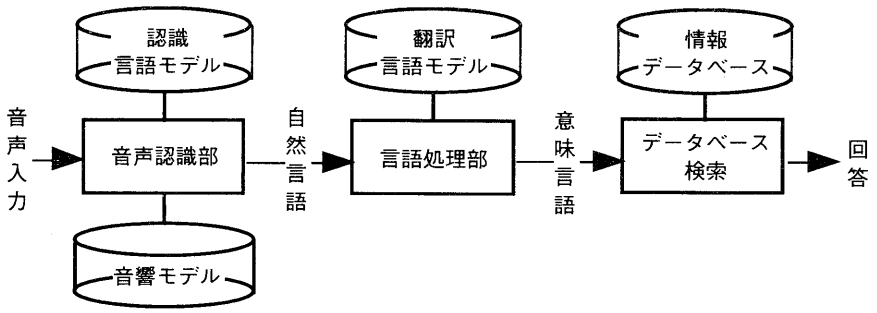

図1　音声理解システムの一構成例

数をフランス語58000語、英語41000語とし、170万対の並行テキストを用いて翻訳言語モデルを推定し、オープンな100文で評価を行い、60文は受容できる翻訳だったと報告している[5]。大量のテキストを用いてはいるが、生テキストから直接翻訳言語モデル $P(\mathrm{e})P(\mathrm{f}\mid\mathrm{e})$ を推定しているため、推定すべきパラメータ空間がスパースで、効率的に翻訳言語モデルが推定できていないと思われる。

Pieraccinらは音声理解のために、意味構造を統計モデルを用いて抽出する方法を提案している[7,8]。この方法では、音声認識の結果としての単語列を統計的モデルを用いて概念ラベルの系列に変換する。すなわち、音声理解の問題を

$$P(\hat{W},\hat{C}\mid A)=\max_{W\times C}P(W,C\mid A) \qquad (3)$$

ただし、A:観測音響信号、W:単語列、C:概念ラベル列

を満たす $\hat{W}$、$\hat{C}$ を求める問題として定式化する。

$$P(W,C\mid A)=\frac{P(A\mid W,C)P(W\mid C)P(C)}{P(A)} \qquad (4)$$

であるから、$P(A\mid W,C)\approx P(A\mid W)$ を音響モデル、$P(W\mid C)P(C)$ を概念言語モデルとしてそれぞれHMMとして推定する。概念言語モデルは、単語列が観測値、概念ラベル系列が観測されない（hiddenな）状態系列に対応するHMMである。

Vidalらは、Pieraccinの枠組みにおける概念を文法規則に置き換えて、Brownの翻訳言語モデルと同様の考え方を、音声理解のための翻訳言語モデルに適用する方法を提案している[9]。Vidalらの方法ではまず自然言語の文法規則、意味言語の文法規則をError Correcting Grammar Inferenceアルゴリズム[10]を用いて学習テキストから推定し、それら文法規則間の条件付き確率を翻訳言語モデルとしている。Vidalらの実験ではATISコーパスのうち比較的翻訳が容易なサブセットを選択して評価を行い、90%程度の文が正確に意味言語に翻訳できたと報告している。しかし、実際には入力される自然言語の表現のバリエーションは非常に多く、したがって文法規則も非常に多くなってしまう。文法規則数が多いと条件付き確率の推定が困難になる点が問題となる。

以下では、入力される自然言語のバリエーションが多い場合にも有効な言語モデルが推定可能とする方法を提案する。本方法は、McCandlessらによる文脈自由文法の推定法[11,12]を用いて、有限状態オートマトンで表現された文法ネットワークの状態数を削減することにより、翻訳言語モデルの推定精度を向上する。これにより入力自然言語の表現のバリエーションをカバーしながら、簡潔な文法規則を構築することができる。ARPAのATISタスクを対象として評価を行い提案法の有効性を示す。

4. 音声理解のための言語処理

図2に統計的言語モデルを用いた音声理解の処理の流れを示す。まず、上段は翻訳言語モデルを推定する過程である。入力自然言語、出力意味言語のそれぞれの学習テキストセットを用いてECGIアルゴリズムにより文法ネットワークを生成する。自然言語文と意味言語文の各対について文法規則列（すなわち、状態

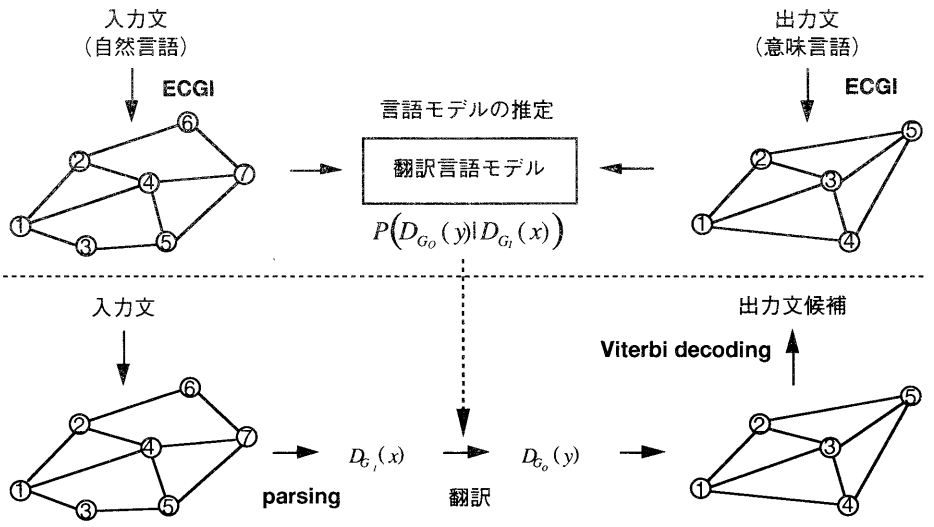

図2　統計的言語モデルによる音声理解の言語処理

遷移系列）を求め文法規則が適用される条件付き確率を推定する。下段の翻訳処理においては、入力文をまず文法ネットワークによりパージングして得られる文法規則系列$D_{G_I}(x)$に対して、先に求められた文法規則の条件付き確率$P\left(D_{G_O}(y) \mid D_{G_I}(x)\right)$を翻訳言語モデルとして適用し、出力側の文法規則系列$D_{G_O}(y)$を求める。この$D_{G_O}(y)$に対応する最適パスをViterbi探索により出力文法ネットワーク内で求めれば、出力文候補を決定できる。

4.1 文法ネットワークの生成

翻訳言語モデルを推定するには、まず自然言語（入力言語）、意味言語（出力言語）各々についてECGI(Errror Correcting Grammar Inference)アルゴリズム[10]により有限状態オートマトンで表現される文法ネットワークを生成する。ECGIアルゴリズムは、学習セット中のテキストを一文ずつパージングし、その文を受理するために必要な状態、遷移を付加していく。入力文と文法ネットワーク中のパスの最適アライメントをError Correcting Parsing（ECP）と呼ばれるDP的手法により求め、その最適パスに沿って必要な状態、遷移を付加する。入力文Xと文法ネットワークにより生成される文Yとの最適アライメントを求めるためのECPの距離尺度には、次式で定義されるLevensteinの提案による距離を用いる。

$$D(X,Y)=\min_{s}(p\cdot sub_s+q\cdot ins_s+r\cdot del_s) \qquad (5)$$

ここで、p、q、rは重み係数で、sは文法ネットワークにおける状態系列、sub_sは置換誤り、ins_sは挿入誤り、del_sは脱落誤りである。実験では$p=q=r=1$とした。図3は、(aaabbccc, aabaccc, aabcccc, aaabccc, aaabacc)という学習テキストを与えられた場合にECGIアルゴリズムで学習されるネットワークの例である。

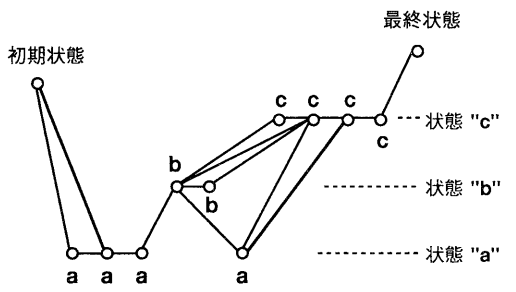

図3 ECGIアルゴリズムで生成される文法ネットワークの例

ECGIアルゴリズムによって推定される文法の状態数は、学習セット中の文の提示順序により異なる。すなわち、単語数の多い文から提示すれば、単語数の少ない文はすでに有限状態オートマトンに存在する状態間を遷移することになり、その結果状態数の少ない文法が生成される。

4.2 翻訳言語モデルの推定

次に入力言語の文法規則と出力言語の文法規則の条件付き確率を、入出力文が一対になった学習テキストにおける共起頻度から推定する。

入力言語の文法をG_I、出力言語の文法をG_Oとすると、与えられた問題は入力文xに対して次式を満足する出力文$\hat{y}$を求めることとなる。

$$\hat{y}=\arg\max_{y\in L(G_o)} P(y\mid x) \qquad (6a)$$

$$=\arg\max_{y\in L(G_o)} P(x\mid y)P(y) \qquad (6b)$$

文章x、yはそれぞれ文法規則の系列$D_{G_i}(x)$、$D_{G_o}(y)$として表現できる。

$$D_{G_i}(x)=\left\{r_{I1}^{x},r_{I2}^{x},...,r_{In}^{x}\mid r_{Ii}^{x}\in G_I\right\} \qquad (7)$$

$$D_{G_o}(y)=\left\{r_{O1}^{y},r_{O2}^{y},...,r_{Om}^{y}\mid r_{Oi}^{y}\in G_O\right\} \qquad (8)$$

G_I、G_Oが正規文法ならば$D_{G_i}(x)$、$D_{G_o}(y)$は一意に決定できる。文法にあいまい性がある場合にはViterbiアルゴリズムによって近似的に求めることができる。これより、(6)式は(9)式のように書き直せる。

$$\hat{y}=\arg\max_{y\in L(G_o)} P\left(D_{G_o}(y)\mid D_{G_I}(x)\right) \qquad (9a)$$

$$=\arg\max_{y\in L(G_o)} P\left(D_{G_I}(x)\mid D_{G_o}(y)\right)P(y) \qquad (9b)$$

Vidalらは、

$$P\left(D_{G_I}(x)\mid D_{G_o}(y)\right) \approx \prod_{r_O\in D_{G_O}(y)}\left(\prod_{r_I\in D_{G_I}(x)} P(r_I\mid r_O)\prod_{r_I\notin D_{G_I}(x)} P(not\ \ r_I\mid r_O)\right) \qquad (10)$$

として(9b)式に基づく推定を試みている。$P\left(D_{G_o}(y)\mid D_{G_i}(x)\right)$でなく$P\left(D_{G_i}(x)\mid D_{G_o}(y)\right)P(y)$を推定しているのはBrownらの翻訳モデルの考え方を継承していると考えられる。(10)式では入力文法規則r_Iが出力文法規則r_Oと共起しない確率$P(not\ \ r_I\mid r_O)$を用いている。文法規模が非常に小さい場合には$P(not\ \ r_I\mid r_O)$

が有効な場合も考えられるが、現実的なタスクを扱う文法の場合には、本質的に無関係な文法規則間の共起頻度を有意な共起頻度と同等に評価することによる悪影響があると考えられる。計算上も項数の増加により確率の積が非常に小さくなるなどの悪影響がある。我々は、出力文法の方が入力文法より規模が小さいことより、(9a)式に基づいて、$P\left(D_{G_I}(x) \mid D_{G_O}(y)\right)P(y)$ではなく$P\left(D_{G_O}(y) \mid D_{G_I}(x)\right)$を推定する方が有利と考え、

$$P\left(D_{G_O}(y) \mid D_{G_I}(x)\right) \approx P\left(r_O \mid r_{I1}^x, r_{I2}^x, \ldots, r_{In}^x\right) \approx \frac{N(r_O, x)}{N(x)} \quad (11)$$

として推定することとした。$N(x)$はG_Iが生成可能な文全体の数であり実際には計算不可能なので以下のように近似を行う。

$$\hat{P}\left(r_O \mid r_{I1}^x, r_{I2}^x, \ldots, r_{In}^x\right) \approx \frac{P^\alpha(r_O) \prod_{r_I \in D_{G_I}(x)} P^\beta(r_O \mid r_I)}{\sum_{r \text{ with the same initial state as } r_O} \left[P^\alpha(r) \prod_{r_I \in D_{G_I}(x)} P^\beta(r \mid r_I) \right]} \quad (12)$$

以上に述べた翻訳言語モデルの推定では、文法G_I、G_Oの規模が大きいほど、つまり、文法規則数が多いほど、$P\left(D_{G_O}(y) \mid D_{G_I}(x)\right)$の推定がスパースになるという問題がある。したがって文法規則数が少ないことが望ましい。文法規模を縮小するため、次節に述べるような方法で有限状態オートマトンの状態数を削減した。

4.3 文脈自由文法の推定による文法状態数の削減

McCandlessらはコーパスから確率的文脈自由文法を推定する方法を提案している[11,12]。この方法では単語bigram確率を距離尺度として、ボトムアップに文法規則を生成する。単語や非終端記号間の距離は(13)～(15)式のように定義する。

$$\left\| u_i, u_j \right\| = d(P_i, P_j) + d(P_j, P_i) \quad (13)$$

$$d(P_i, P_j) = \sum_{C \in Context} P_i(C) \times \log \frac{P_i(C)}{P_j(C)} \quad (14)$$

$$\begin{aligned} P_i(C) &= P(C \mid u_i) \\ &\approx P(u_{left}, u_i \mid u_i) \times P(u_i, u_{right} \mid u_i) \\ &\approx \frac{N(u_{left}, u_i)}{N(u_i)} \times \frac{N(u_i, u_{right})}{N(u_i)} \end{aligned} \quad (15)$$

すなわち、同じ文脈で出現する頻度の高い単語同士の距離を近いと判断し、マージして新たな非終端記号を生成する。単語がマージされて非終端記号となることで、有限状態オートマトンの状態数を削減することができる。

4.4 タスクに依存しない知識の利用

タスク固有の知識を必要としない前処理によっても文法上の状態数の削減が可能である。例えば、Boston、New Yorkなどの単語は、「地名」という一つの非終端記号にまとめることができる。同様に、数字、日付、曜日、月、航空会社、空港名、座席クラスなども、具体的な単語の代わりに非終端記号で置き換え可能である。

文脈自由文法による非終端記号へのマージと、この前処理により、入力言語の文法状態数は約2500から約1500へ、出力言語の状態数は約1000から約600へ削減できた。

5. 実験

LDC(Linguistic Data Consortium)より頒布されているATIS2のコーパスより文脈に独立にWIN文を生成可能なクラスAに分類されたデータのうち、WIN文に括弧の含まれないデータを対象として実験を行った。1915文を学習に、213文を評価に用いた。

今回の実験では、言語処理部のみの評価を行うこととしたため入力は音声認識誤りの含まれていないテキストである。

5.1 文法ネットワークの生成

ECGIアルゴリズムによって生成された文法ネットワークの評価セットに対するカバー率、すなわち、正しくパージングできた文の割合は約90%であった。

ECGIアルゴリズムでは一文ずつ逐次的に評価しながら状態／遷移を付加していくため、文を与える順序によって結果が異なる。文の提示順序を、単語数で数えて昇順、降順、コーパス中にあった通りの順とした場合に生成される文法ネットワークの状態数、遷移数がどのようになるか検討した。結果を表1に示す。降順で提示したものは、コーパス中にあった順で提示するのに比較して状態数で23%、遷移数で10%規模が小

表1　ECGIアルゴリズムにおける文提示順の効果

	状態数	遷移数
降順	1206	3184
ソートせず	1489	3505
昇順	2181	3968

さくなっている。

この結果が示すようにアルゴリズムの評価基準が、学習セット中のすべての文を同時に評価するグローバルな尺度になっていないため、同じ非終端記号が、生成された文法ネットワークの異なる場所にいくつも存在するなど、冗長性があることが明らかになった。この部分については、今後マルコフモデル、HMMなど学習セット全体を同時に評価する基準で学習が進められる方法にしていく必要があると思われる。

5.2　自然言語から意味言語への翻訳

評価は翻訳結果と正解WIN文との文章単位での正解率により行った。(12)式のβは1-αとして、αを実験的に最適化して実験を行った結果、学習セットに対する文正解率は95.3%、評価セットに対しては62.4%であった。評価セットに対する性能はまだ十分とは言えない。ARPAにおける他の評価結果(unweighted error rate: 3.8%～30.6%)と比較するため、置換、脱落、挿入を考慮した単語誤り率(unweighted error rate)を求めたところ、14.2%であった。この結果は、コーパスからの自動推定を重視して、タスク固有のヒューリスティックスなどを用いていないことを考慮すれば、有望な結果と考えられる。

6. まとめ

音声理解のために、自然言語を意味言語に翻訳する言語モデルをコーパスから自動的に推定する方法について述べた。現実的な問題を対象とする場合には、文法の規模が大きくなり、それにともない翻訳言語モデルの推定が難しくなるため、文法上の状態数を削減して、翻訳言語モデルの推定精度を向上する方法を提案した。ATIS2のコーパスにより評価を行い、コーパスから翻訳言語モデルを自動獲得することの可能性を示した。ECGIアルゴリズムにより推定される文法の冗長性や、文脈自由文法の推定において同一コンテキストでの出現頻度のみで単語をマージしていることなどが、問題として考えられる。今後これらの問題点を解決するとともに、導入が容易なものについては、タスク固有のヒューリスティックスも考慮しながら、音声入力から意味理解までの全体的な評価を行っていきたい。

謝辞

統計的機械翻訳の関連研究をご教示下さった東京工業大学田中穂積教授、統計的機械翻訳の関連論文を送ってくださったNTT情報通信研究所永田昌明氏、ECGIアルゴリズムのプログラムを提供してくださったUniv. Politecnica de ValenciaのVidal教授、修士論文を送ってくださったMITのMcCandless氏に感謝します。

参考文献

1. W. Chou, T. Matsuoka, B. H. Juang, and C. H. Lee, "An algorithm of high resolution and efficient multiple string hypothetization for continuous speech recognition using inter-word models," Proc. ICASSP-94, pp.II-153-156
2. 永田、技術早分かり「統計手法を利用した機械翻訳」、AMMTジャーナル No. 7、pp. 36-39、アジア太平洋機械翻訳協会、1994
3. 永田、自然言語処理と学習理論、言語処理学会第1回年次大会チュートリアル資料、pp. 1-20、1995
4. P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," Computational Linguistics, Vol. 16, No. 2, pp. 79-85, 1990
5. P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. D. Lafferty, and R. M. Mercer, "Analysis, Statistical Transfer, and Synthesis in Machine Translation," Proc. International Conference on Theoretical Methodological Issues in Machine Translation, pp. 83-100, 1992
6. P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," Computational Linguistics, Vol. 19, No. 2, pp. 263-311, 1993
7. R. Pieraccini and E. Levin, "Stochastic Representation of Semantic Structure for Speech Understanding," EUROSPEECH-91, Proc. Vol. 2, pp. 383-386, 1991
8. R. Pieraccini, E. Levin, and E. Vidal, "Learning How to Understand Language," EUROSPEECH-93
9. E. Vidal, R. Pieraccini, and E. Levin, "Learning associations between grammars: a new approach to natural language understanding," Proc. EUROSPEECH-93, pp. 1187-1190
10. H. Rulot, N. Pietro, and E. Vidal, "Learning Accurate Finite-State Structual Models of Words through the ECGI Algorithm," Proc. ICASSP-89, pp. 643-646, 1989
11. M. K. McCandless and J. Glass, "Empirical acquisition of word and phrase classes in the ATIS domain," Proc. EUROSPEECH-93, pp. 981-984
12. M. K. McCandless, "Automatic Acquisition of Language Models for Speech Recognition, " Thesis for M.E., Dept. of Electrical Engineering and Computer Science, MIT, June, 1994